# TECHNICAL REVIEW: PERFORMANCE OF EXISTING IMPUTATION METHODS FOR MISSING DATA IN SVM ENSEMBLE CREATION

Shahid Ali and Simon Dacey

Department of Computing, Unitec Institute of Technology, Auckland, New Zealand

***ABSTRACT***

*Incomplete data is present in many study contents. This incomplete or uncollected data information is named as missing data (values), and considered as vital problem for various researchers. Even this missing data problem is faced more in air pollution monitoring stations, where data is collected from multiple monitoring stations widespread across various locations. In literature, various imputation methods for missing data are proposed, however, in this research we considered only existing imputation methods for missing data and recorded their performance in ensemble creation. The five existing imputation methods for missing data deployed in this research are series mean method, mean of nearby points, median of nearby points, linear trend at a point and linear interpolation respectively. Series mean (SM) method demonstrated comparatively better to other imputation methods with least mean absolute error and better performance accuracy for SVM ensemble creation on CO data set using bagging and boosting algorithms.*

***KEYWORDS***

*Ensemble, Bagging, Boosting, Imputation, Aggregation*

## 1. INTRODUCTION

Air quality is monitored to detect any pollutant concentrations that has adverse effects on human beings [1]. For this air quality is monitored at various locations through various monitoring stations. However, to conduct air pollution analysis which has large observations of missing data makes the task difficult to evaluate [13]. The missing data is a result of equipment failure, human error, routine maintenance, changes in sitting of monitors or due to some other factors [19]. This missing data or incomplete data set creates results that are different from those that would have been monitored through complete data set [15].

The occurrence of missing data requires a serious consideration on analysing the data. In fact, there are three main problems associated in dealing with incomplete data set [33] [23]. Firstly, the loss of missing information results in reduction of efficiency. Secondly, the missing data leads to problems in data handling, computation analysis and further minimizing the efforts to use the standard software. Thirdly, which is the most important, the results produced via missing data may be biased due to the difference between the observed data and unobserved data. Currently, there are some statistical packages such as SPSS which can handle missing data and can perform replacement for missing values.

Our approach to handle missing values or incomplete data set in current research is limited to five imputation options. These are implemented in SPSS and our goal for this research is to determine the best imputation method to replace missing values for Carbon monoxide(CO) concentrations in our research. These five imputation methods for missing data are explained later in this chapter.

The remainder of this research is organized as follows: Section 2 provides the types of missing data. Section 3 provides a literature review on previous computational studies for missing data. Section 4 discusses the reasons for using imputation methods for missing data. Section discuses about SVM ensemble for air pollution data analysis and methods for constructing SVM ensemble.

DOI: 10.5121/ijdkp.2017.7606 

Section 6 provides experimental design and imputation methods used to handle missing data. Section 7 is dedicated towards results and discussion for the experiments. Finally, section 8 presents conclusion to this research.

## 2. TYPES OF MISSING DATA

Incomplete data is present in many study contents [28]. This incomplete or uncollected data information is named as missing data (values), and considered as vital problem for various researchers. Even this missing data problem is faced more in air pollution monitoring stations [16], where data is collected from multiple monitoring stations which are widespread across various locations [3]. Generally, there are two types of missing data encountered in air quality monitoring [30]. The first form of missing data is non-ignorable data, where missing datum probability is dependent on its value, and ignoring missing data probability of missing datum does not rely on its value. The second form of missing data is ignorable missing data, which is of two types. The first type of ignorable missing data is linked to sampling, which refers to the situations where it is not possible to obtain data from whole population. In this case probability sampling is used to get a representative population sample. The second type of ignorable missing data is where data is missing at random (MAR), it refers to the pattern of missing that vary for subsets for a variable. It is determined that the air quality data referred to MAR.
To test the accuracy of imputation method, from a complete data set incomplete data sets need to be generated [32]. For the imputation of missing values of air quality, various patterns of air quality missing data sets are created to evaluate the efficiency of each method. These missing patterns helped researchers to select the best estimation imputation method for research analysis.

## 3. LITERATURE REVIEW

Missing data is a serious problem, that creates uncertainty in research results [20]. In literature, various methods and techniques are proposed to address imputation of missing data which will be discussed here briefly.

Mean top bottom technique was applied to replace the missing values in $PM_{10}$ concentrations in a data set [32]. It was found in the research that this method performed very well only when the missing data was in small number.

Nearest neighbor method was proposed for the imputation of incomplete $PM_{10}$ concentration data [13]. Further in this study three other methods namely, mean substitution, expectation maximisation (EM) and hot deck were also considered for imputation of missing data.

Mean, median, hot deck, KNN and mean method by step depression imputation methods were used to improve the imputation accuracy of each method through well know classifiers KNN, SVAR, SVMP, C4.5, RIPPER and LSVM [25]. Statistical results of this study shown that mean method by step depression (MMSD) results were more acceptable compare to other methods and resulted in better performance of the classifier with missing values of 7.72% to 20% [25].

Traffic control, traffic management and control applications require complete and accurate data because of various reasons, however, such data is sometime unavailable [34]. For this typical problem researchers categorized the imputation methods into three categories i.e., prediction methods, statistical methods and interpolation methods. Results from various studies demonstrated that statistical methods were effective in imputing missing data resulted in better performance results and low reconstruction errors [17]. A similar study for traffic flow missing data with ten methods was conducted [7]. The performance of those methods was compared with Bayesian Principal Component Analysis (BPCA) imputation methods [7]. Experiment analysis outperformed the results of BPCA imputation methods and demonstrated good choices in dealing with missing data.

Incomplete data plays important role in prediction accuracy, as the incomplete data is present both in training and testing data set tends to produce biased results [27]. It is quite evident from various researches that combining the output of various classifiers results in the prediction accuracy [2]. In this regard two ensemble based imputation techniques namely, Bayesian multiple imputation and nearest neighbor single imputation [26] for imputation of missing data were proposed. Results of this study demonstrated better results with decision trees support method.

Environmental monitors, scientific researchers and process controllers have widely used time series data for analysis. However, in the presence of missing data time series results enforce big question mark. In this regard to address the time series missing data imputation method based on Genetic Programming (GP) and Lagrange Interpolation was proposed [9]. The results of this study were promising and produce efficient results on imputation missing data in time series and further possessed no loss to data sets statistical properties leading to better understanding of missing data pattern.

From the previous literature, it is quite evident that various methods based on machine learning for imputation of missing data were proposed. However, for our research we will take a different approach for imputation of missing data of Carbon monoxide (CO) concentrations in Auckland region for air pollution analysis by deploying series mean method, mean nearby point method, median nearby point method, linear interpolation method and linear trend at point method. For all above methods mean absolute error will be calculated and each method classification accuracy will be determined by SVM ensemble creation. The above imputation methods are explained further in this research.

## 4. REASONS FOR USING IMPUTATION METHODS

Researchers have used various alternative methods for imputation of missing data [22]. The missing data is in various researches is handled by three traditional reasons:

1. The computer programs are defined in such a way that an empty space is a missing value. Therefore, computer programs ignore these missing values as defined, in other words they do not include them in the analyses [11].

2. Another common method to interfere into missing values is to remove the variable or subjects for which missing values are there [14]. However, deleting the subjects may result in loss of data and will produce biased results, because of the systematic difference between the collected and uncollected data [21].
   On the other side if the missing values are presented in a group of variables, then if the variable(s) is/are of no such importance then the variable can be deleted. However, in case where the available variables are distributed, then deleting such variable(s) will be of serious loss of data [21]. Moreover, variables who have missing values are not distributed randomly, then deleting those variable data may result in skewness of the distribution [8]. For such reasons, it is proposed that imputation of missing values helps to protect the sample size as well [28].

3. Another way to solve missing data problem is to make predictions of missing values and use them in the analysis [5] [6]. However, prediction of missing values and imputation can only be used for quantitative variables. The three most common methods for predictions of missing quantitative variables [8] [24], are prior knowledge, regression and average (mean) imputation.

## 5. SVM ENSEMBLE

Computational air pollution data analysis is spatio-temporal in nature [2], this research focuses on constructing dynamic computing environment through SVM ensemble. Various individual SVMs

are aggregated for the purpose of data mining, where SVM aggregation results in outstanding generalizability and speedy parallel computation [1][2].

$$X = \begin{bmatrix} X_{t1} \\ X_{t2} \\ \vdots \\ X_{tn} \end{bmatrix} \quad (1)$$

The air pollution dataset can be represented as (1) and (2), which is a three dimensional matrix. It can be further simplified as time series of two dimensional data matrix (1), as environmental data is gathered over time line. Similarly, in time series one time instance is a matrix (2). Air pollution states are represented by elements in various geometric location, where Elements are represented as data and are collected by various sensor devices in different locations.

$$X_{ti} = \begin{bmatrix} x_{1,1,t_i}, & \cdots, & x_{1,m,t_i} \\ \vdots & \ddots & \vdots \\ x_{n,1,t_i}, & \cdots, & x_{n,m,t_i} \end{bmatrix} \quad (2)$$

The individual SVMs decisions are aggregated by majority of vote method to analyse the air pollution problem.

### 5.1 METHODS OF SVM ENSEMBLE CONSTRUCTION

Bagging and boosting algorithms are used for the construction of SVM ensemble and the imputation methods are evaluated based on that.

#### 5.1.1 Bagging

Bagging algorithm generates various bootstrap training sets from the original training set and deploys each of them to produce a classifier for the enclosure in ensemble. The bagging algorithm and bootstrap sampling with replacement is illustrated below [36].

BAGGING(T,M)

1 For each $m = 1,2,...,M$,
2 $T_m = Sample\ With\ Replacement(T,N)$
3 $h_m = L_b(T_m)$
4 Return $h_{fin}(x) = argmax_{y \in Y} \sum_{m=1}^{M} I(h_m(x)) = y$

SAMPLE WITH REPLACEMENT(T,N)

1 $S = \phi$
2 **for** $i = 1,2,...,N$
3 $r = randominteger(1,N)$ 4 Add $T[r]$ to $S$
5 Return S.

In order to create a bootstrap sample from a training set of $N$, we execute N multinomial trials and in each trial we draw one of the $N$ samples. In this case each sample has a probability of $1/N$ to be drawn in each trial.

The second algorithm shown above exactly does this $N$ times, the algorithm selects a number from 1 to $N$ and then adds the $r^th$ training example, to bootstrap training set $S$. Noticeably, some

of the original training examples will not be selected for inclusion of bootstrap training set and others will be selected one time or more. In bagging, the number of base classifiers that need to be learned $M$, are created through bootstrap training sets and further classifiers are generated using each of them. Bagging yields a function $h(x)$ that classifies new examples by yielding the class y that receives the maximum number of votes from the base models $\{h_1,h_2,h_3 ... h_m\}$. In bagging, the $M$ bootstrap training sets produced are likely to have some differences. If these differences are enough to show obvious differences among the $M$ base models, then in that case the ensemble will perform better than the base models individually [35] [12].

Models are said to be unstable [18], if the differences in their training sets show significant differences in the models and stable if not. In other way, we can say that bagging method does more to reduce the variance in base models instead of bias. So, bagging performs better relative to its base models, when the base models have low bias and high variance.

### 5.1.2 BOOSTING

Adaboost is a boosting algorithm which we used with other algorithms for spatial and temporal air pollution analysis in our research. Adaboost algorithm generates a sequence of based models along with different weight distributions over training set. Adaboost algorithm is illustrated below [36].

ADABOOST$^{(}$ $\{(x_1,y_1),...,(x_N,y_N)\},L_b,M)$

1 Initialize $D_1(n) = 1/N$ for all $n \in \{1,2,...,N\}$
2 **for** $m = 1,2,...,M,$
3 $h_m = L_b(\{(x_1,y_1),...,(x_N,y_N)\},D_m))$
4 Calculate the error of$\geq$ $h_m : \epsilon_m = \sum n_{:hm(xn)=/yn} D_m(n)$
5 If $\epsilon_m$ 1/2 then
6 set $M = m - 1$ and abort this loop
7 Update distribution $D_m$:
8 $Dm+1(n) = Dm(n) \times \{$ $\frac{1}{\epsilon}$ $2(1\underline{1}_{m-} m_{)} m(xn) = yn$ if $h$
$2\epsilon$ otherwise
9 Output the final hypothesis:
10 $fin(x) = argmax_{y\in Y} \sum_{m:h_m(x)=y}$ $\frac{1-m}{\epsilon_m} h$ $log$ $\epsilon$

It has a set of $N$ training examples, a base model learning algorithm $L_b$ and the number of base models $M$, that we want to combine. Adaboost algorithm was designed for two class classification. However, it is regularly used in previous researches for more than two classes.
The first step in Adaboost algorithm is the construction of weights distribution $D_1$ over the training set. In Adaboost algorithm the first distribution is one that assigns equal weight to all $N$ training examples. By now, we enter into the loop of the Adaboost algorithm. In order to make first base model, we call the base model learning algorithm $L_b$ with distribution $D_1$ over the training set. Failure of $L_b$ to take weighted training set, one can derive it by sampling with replacement from the original training set with the help of distribution $D_m$. After getting $h_1$ hypothesis and calculating error $E_1$ on the training set, which is the sum of the weights of the training examples that $h_1$ misclassifies.

We want $E_1 < 1/2$, if this condition is not satisfied then we stop here and will to ensemble that consists previously generated base models. In this case if $\epsilon_1 < 1/2$ is satisfied, then we calculate $D_2$ over the training examples as follows. Correctly classified examples by $h_1$ have their weights multiplied by $\frac{1}{2(1-\epsilon_1)}$. Misclassified examples by $h_1$ their weights will be multiplied by $\frac{1}{2(\epsilon_1)}$. According to our condition $\epsilon_1 < 1/2$, the weights of correctly classified examples will be reduced and the weights of misclassified examples will be increased [36]. In other words, examples that $h_1$ misclassified their aggregate weight will increase to 1/2 under $D_2$ and examples that $h_1$ correctly

classified their aggregate weight will reduce to 1/2 under $D_2$. From here we go into the next iteration of the loop to construct base model $h_2$ using training the set and new distribution $D_2$. We build $M$ based models in this way. The ensemble derived from Adaboost is a function that takes new example as an input and returns the class that gets the maximum weighted vote over the $M$ base models.

Each base model's weight is $log(\frac{1-\epsilon_m}{\epsilon_m})$, which is proportional to accuracy of base model on the weighted training set presented on it.

It is quite clear from the above explanation that the core of Adaboost algorithm is the distribution updating step. In the Adaboost algorithm we perceive that $\epsilon_m$ represents the sum of the weights of misclassified examples. The weights of misclassified examples are multiplied by $\frac{1}{2\epsilon_m}$, by doing this, sum of their weights increased by

$$\epsilon_m * \frac{1}{2\epsilon_m} = \frac{1}{2}.$$

Correctly classified examples weight is $(1 - \epsilon_m)$ but their weights are multiplied by $\frac{1}{2\epsilon_m}$ , hence, sum of their weights decreases by

$$\left(1-\epsilon_m\right) * \frac{1}{2\left(1-\epsilon_m\right)} = \frac{1}{2}.$$

The adjustment of this weight results in the next model is to be generated by weak learner, which will have an error less than 1/2. From this misclassified example of previous base model will be learned.

In general boosting algorithm reduces the bias than variance. For this boosting algorithm tends to improve its base models when they have high bias and low variance. The reduction of bias in boosting algorithm derives from the fact that it adjusts distribution over the training set. The weights of misclassified examples by base model increases, resulting in base model algorithm to focus more on those examples. In an instance, when the base model learning algorithm is biased to certain examples gets more weight resulting the possibility of correcting that bias. This mechanism of adjusting the training set distribution results in difficulty for boosting algorithm, when the training data is noisy [37]. Noisy examples are difficult to operate and learn in boosting algorithm [37]. Because high weights are assigned to noisy examples compare to others, causing boosting algorithm to focus more on those examples and overfit the data.

## 6. EXPERIMENTAL DESIGN AND IMPUTATION METHODS

For simulation of missing data an annual hourly monitoring records for CO concentrations is collected from seven stations in Auckland region, Takapuna, Khyber Pass road, Henderson, Pakuranga, Queen Street, Glen Eden and Pukekohe. The data set contains CO concentrations on a time scale of one per hour (hourly averaged) spread over a year.

For calculation of missing values of CO concentrations of seven monitoring stations and for the calculation of mean absolute error of each imputation method we use IBM SPSS Statistics version 22 for our experiments. Whereas, for the classification accuracy of each imputation method we run Matlab on Windows 7 Enterprise with system configuration Intel Core i5 processor (3.2 Ghz) with 4 GB 1067 MHz DDR3of RAM.

Characteristics of CO are shown in table Table 1. Table 1 shows that a total of 8783 observations of CO are available for experiment purposes of which 2169 (24.69 %) are missing. Number of extremes of seven monitoring stations is provided, which was varies from station to station.

Table 1. Characteristics of CO data

| Stations | N | Mean | Std. | Missing | | No. of Extremes | |
|---|---|---|---|---|---|---|---|
| | | | | Count | Percent | Low | high |
| Station 1 | 8544 | .427 | .5505 | 239 | 2.7 | 0 | 332 |
| Station 2 | 8656 | 1.212 | 1.1644 | 127 | 1.4 | 0 | 412 |
| Station 3 | 8487 | .256 | 3484 | 296 | 3.4 | 0 | 399 |
| Station 4 | 8504 | .501 | .6049 | 279 | 3.2 | 0 | 351 |
| Station 5 | 8477 | .695 | .6247 | 306 | 3.5 | 0 | 349 |
| Station 6 | 8564 | .310 | .4121 | 219 | 2.5 | 0 | 412 |
| Station 7 | 8080 | .254 | .3083 | 703 | 8.0 | 0 | 309 |

Figure 1 is an illustration of concentrations of CO data skewness. Figure 1 shows that there is some variability in range as shown in concentrations of CO data from 3.9 to 8.9 $\mu g/m^3$ of various monitoring stations. Whereas, the data is skewed towards the right demonstrating most of the time low concentrations of CO were observed across Auckland region. Environmental Performance Indicators (EPIs) are calculated relative to the National Environmental standards for Air Quality for each gas. Since these values are different the EPI classes are different. The data for this study includes five classes for monitoring CO in Auckland region according to EPIs: (class ”a”: excellent (meaning, air quality is considered fantastic and no risk at all to people), class ”b”: good (meaning, air quality is considered satisfactory and there is little to people health), class ”c”: acceptable (meaning, air quality is acceptable, however, there is risk to people health), class ”d”: alert (meaning, air quality is not acceptable and there is serious risk to people health), class ”e”: action (meaning, air quality is deteriorating and a quick response is required).

Hence to deal with the missing data of CO on annual hourly monitoring records of various stations in Auckland region requires a method(s) for imputation of missing data.

For this analysis in our experiments for missing data of seven monitoring stations of CO concentrations we applied five imputation methods that are implemented in SPSSM. These methods are named as series mean (SM) method, mean of nearby points (MNP), median of nearby points (MDNP), linear trend at a point (LTAP) and linear interpolation (LI). Each method mean absolute error is calculated for seven monitoring stations and based on that its effectiveness determined. Similarly, each imputation method classification accuracy is calculated and further evaluation of each imputation method on performance accuracy using boosting and bagging algorithms is conducted. With the help of each method imputed data ensemble is build and its classification accuracy is computed.

### 6.1 IMPUTATION METHODS

The existing five imputation methods for missing CO data of seven monitoring stations of Auckland region are explained below with their significance.

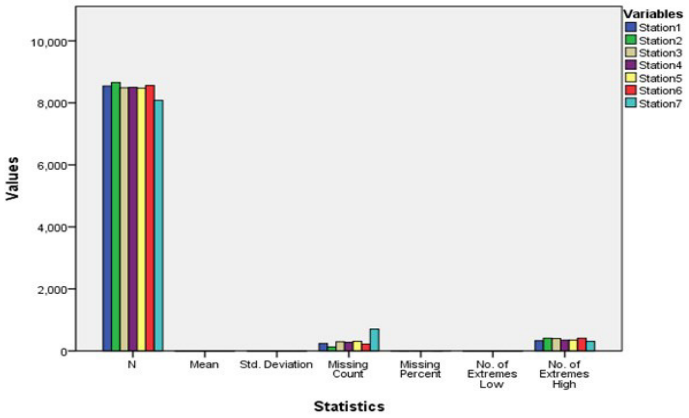


Figure 1. Concentrations of CO data Skewness

### 6.1.1 Series Mean (SM) Method

In this method missing values are imputed with mean of the entire data. The missing CO concentrations of seven monitoring stations were replaced to their station's mean.

### 6.1.2 Mean of Nearby Points (MNP) Method

In this method missing values are imputed by the mean of nearby points (surrounding) values. The number of nearby points is derived from 'span of nearby points' option in SPSS. The default value in the SPSS program is '2 digits'. In other words, the mean is calculated by using complete station's data from above and below missing values, and this value is imputed instead of entire data.

### 6.1.3 Median of Nearby Points (MDNP) Method

In this method missing values are imputed by the median of nearby (surrounding) values. The number nearby points are derived from 'span of nearby points' option in SPSS program. The default value in the SPSS program is '2 digits'. The median is calculated by using the complete values of a station's data from above and below, the missing data and the derived value is used to replace the missing value.

### 6.1.4 Linear Interpolation (LI) Method

This method replaces missing values by interpolation. The last incomplete information in the CO monitoring station's data before the missing value and the first value after the missing data in the CO monitoring station's data are used for interpolation [4]. In case where the first or last data in a series is missing, then the missing value is not replaced.

### 6.1.5 Linear Trend at Point (LTP) Method

Missing values in this method are replaced in accordance with the trend of current structure data. The imputed missing data is replaced based on an index variable scale 1 to n [10].
The performance of each above method is determined based on the mean absolute error (MAE). The selection of best method is based on to estimate the missing values with least error. MAE is the average between actual and predicted data values. It can be represented from (1.1) [13].

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|P_i - O_i| \qquad (1.1)$$

MAE values range from zero to infinity, however, a perfect fit cab only be achieved when MAE=0.

Classification accuracy in building ensemble was another consideration of each method which is considered in evaluation of each method imputation accuracy for missing data. SVM ensemble is developed based on each imputation method using boosting and bagging algorithms. Confusion matrices are obtained for each method in building ensemble for CO analysis.

## 7. RESULTS AND DISCUSSIONS

The descriptive statistics of mean absolute errors with SM method are shown in Table 2. Table 2 shows that with SM method the best result of least MAE of .1533 is obtained for station 7 of CO monitoring. The second least error result is of .2237 for imputing missing data of station 6. Hence, the lower .1533 MAE with SM method shows that prediction imputation of missing data to actual values with this result showed least error when it comes to imputation of missing data.

Table 2. Mean Absolute Errors with SM Method

| Stations | N | Minimum | Maximum | Mean | Std.Deviation |
|---|---|---|---|---|---|
| Station 1 | 8783 | .00 | 6.97 | .2992 | .45308 |
| Station 2 | 8783 | .00 | 7.49 | .9134 | .70834 |
| Station 3 | 8783 | .00 | 3.64 | .2246 | .25858 |
| Station 4 | 8783 | .00 | 7.20 | .3551 | .47763 |
| Station 5 | 8783 | .00 | 8.21 | .4405 | .42733 |
| Station 6 | 8783 | .00 | 5.19 | .2237 | .33987 |
| Station 7 | 8783 | .00 | 6.45 | .1533 | .25284 |

The descriptive statistics of mean absolute errors with MNP method are shown in Table.1.3. Table 3 shows that .165 MAE for station 7 is obtained, this is the best result which is available with this method. The second best result is achieved with a MAE of .228 for station 6. The results of .254 and .256 MAEs are obtained in imputation of predicted actual values. Further results of this method showed how close the results are in terms of MAE for predicting missing values for each monitoring station.

Table 3. Mean Absolute Errors with MNP Method

| Stations | N | Minimum | Maximum | Mean | Std.Deviation |
|---|---|---|---|---|---|
| Station 1 | 8544 | .03 | 6.97 | .3075 | .45657 |
| Station 2 | 8783 | .01 | 7.49 | .9242 | .70189 |
| Station 3 | 8783 | .01 | 3.64 | .2309 | .25621 |
| Station 4 | 8782 | .00 | 7.20 | .3645 | .47509 |
| Station 5 | 8781 | .01 | 8.21 | .4499 | .42179 |
| Station 6 | 8783 | .01 | 5.19 | .2283 | .33967 |
| Station 7 | 8771 | .00 | 6.45 | .1650 | .25036 |

The descriptive statistics of mean absolute errors with MDNP method are shown in Table 4. Table 4 shows that the minimum MAE of .1607 is obtained for station 7. However, MAE of .9236 is highest for station 2. The results of this method demonstrated that least .1607 of MAE is obtained through this method imputation compare to actual values.

Table 4. Mean Absolute Errors with MDNP Method

| Stations | N | Minimum | Maximum | Mean | Std.Deviation |
|---|---|---|---|---|---|
| Station 1 | 8544 | .03 | 6.97 | .3075 | .45657 |
| Station 2 | 8783 | .01 | 7.49 | .9236 | .70298 |
| Station 3 | 8783 | .00 | 3.65 | .2295 | .25715 |
| Station 4 | 8782 | .01 | 7.21 | .3629 | .47628 |
| Station 5 | 8781 | .01 | 8.21 | .4494 | .42276 |
| Station 6 | 8783 | .01 | 5.19 | .2275 | .33991 |
| Station 7 | 8771 | .01 | 6.46 | .1607 | .25324 |

The descriptive statistics of mean absolute errors with LI method are shown in Table 5. Table 5 shows imputation for missing data prediction through Linear Interpolation (LI) received .1620 MAE for station 7. However, the second best result for this method is achieved with a MAE of .2280 for station 6.

Table 5. Mean Absolute Errors with LI Method

| Stations | N | Minimum | Maximum | Mean | Std.Deviation |
|---|---|---|---|---|---|
| Station 1 | 8544 | .04 | 6.96 | .3250 | .44736 |
| Station 2 | 8783 | .01 | 7.49 | .9220 | .70219 |
| Station 3 | 8783 | .00 | 3.65 | .2281 | .25681 |
| Station 4 | 8783 | .00 | 7.20 | .3616 | .47535 |
| Station 5 | 8783 | .01 | 8.20 | .4573 | .41793 |
| Station 6 | 8783 | .00 | 5.19 | .2269 | .33914 |
| Station 7 | 8773 | .00 | 6.45 | .1597 | .25134 |

The descriptive statistics of mean absolute errors with LTP method are shown in Table.1.6. Table.1.6 shows that the best result with LTP is obtained with minimum MAE of .1597 for station 7 for prediction of missing values. Whereas, second best result is obtained with a MAE of .2269 for station 6.

Table 5. Mean Absolute Errors with LTP Method

| Stations | N | Minimum | Maximum | Mean | Std.Deviation |
|---|---|---|---|---|---|
| Station 1 | 8544 | .03 | 6.97 | .3075 | .45657 |
| Station 2 | 8783 | .01 | 7.49 | .9239 | .70339 |
| Station 3 | 8783 | .00 | 3.65 | .2305 | .25769 |
| Station 4 | 8783 | .00 | 7.20 | .3637 | .47602 |
| Station 5 | 8782 | .01 | 8.21 | .4505 | .42326 |
| Station 6 | 8783 | .01 | 5.19 | .2280 | .34028 |
| Station 7 | 8771 | .00 | 6.46 | .1620 | .25268 |

Overall, the SM method demonstrated best in prediction for missing data having lowest MAE of .1533 for station 7. This is followed by the LTP method having MAE of .159 for station 7 also. Relatively all the five imputation methods utilised in this study performed considerably well, however, among the five imputation methods best results are obtained through SM method followed by LTP method with least MAE.

We try to classify the CO data set by using all the above five imputation methods for missing CO data by creating an SVM ensemble with each method missing imputed data. We deployed five imputation methods used in this research for filling missing data in CO analysis, each method classification accuracy was evaluated by creating an ensemble using bagging and boosting algorithms.

Firstly, we deployed SM method for imputation of missing data and created an SVM ensemble with this data. The ensemble obtained with SM method imputed data using adaBoostM1 algorithm resulted in a classification accuracy of 76.9% based on confusion matrix illustrated in Figure 2.

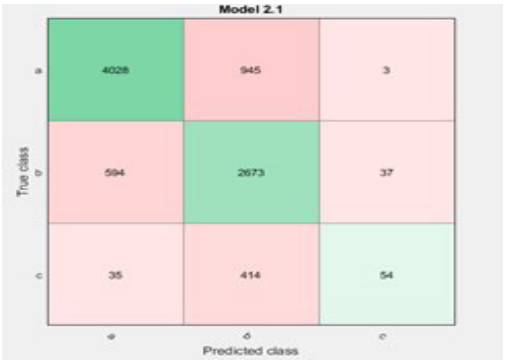


Figure 2. SM method Confusion Matrix AdaBoostM1 Algorithm

Ensemble obtained with SM method using bagging algorithm resulted in 74.6% classification accuracy base on confusion matrix illustrated in Figure 3.

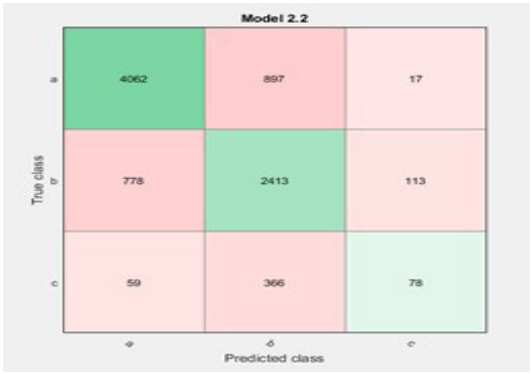


Figure 3. SM method Confusion Matrix Bagging Algorithm

Ensemble obtained with imputed method MDNP with adaBoostM1 algorithm resulted in 76.7% of classification based on confusion matrix as shown in Figure 4. However, ensemble using MDNP method with bagging algorithm resulted in 75.0% classification accuracy based on confusion matrix illustrated in Figure 5.

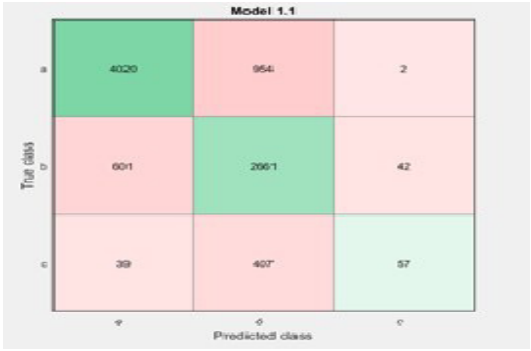


Figure 4. MDNP method Confusion Matrix AdaBoostM1 Algorithm

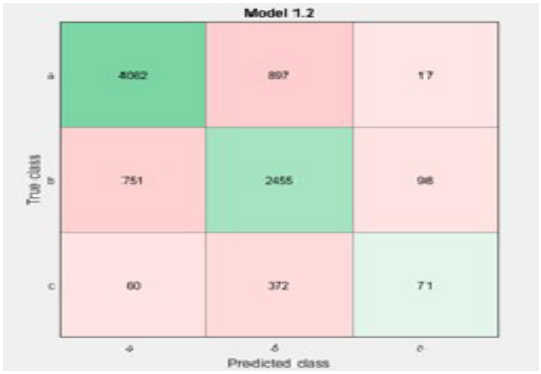


Figure 5. MDNP method Confusion Matrix Bagging Algorithm

Ensemble based on MNP method resulted in 76.7% classification accuracy based on confusion matrix using adaBoostM1 algorithm as shown in Figure 6. With this method classification accuracy of ensemble resulted same i.e. 76.7% using bagging algorithm based on confusion matrix as illustrated in Figure 7.

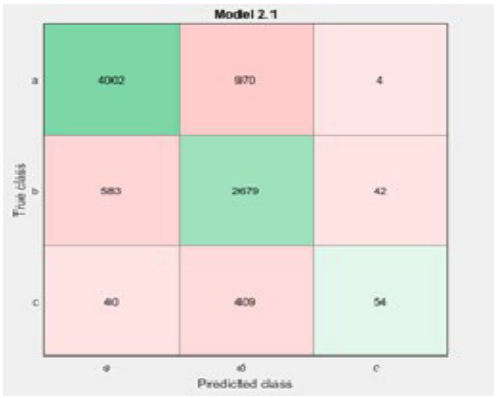


Figure 6. MNP method Confusion Matrix AdaBoostM1 Algorithm


Figure 7. MNP method Confusion Matrix Bagging Algorithm

A 76.9% of classification accuracy based is obtained in ensemble creation with LI method using adaBoostM1 algorithm based on confusion matrix as shown in Figure 8. A similar percentage of 76.9% is obtained using bagging algorithm deploying LI method based on confusion matrix as shown in Figure 9.

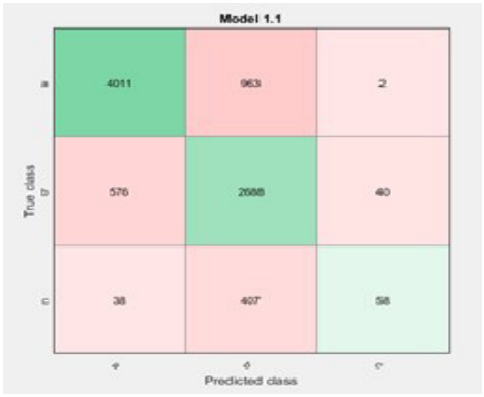


Figure 8. LI method Confusion Matrix AdaboostM1 Algorithm

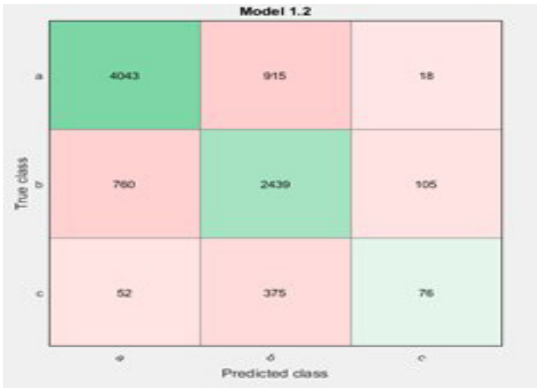


Figure 9. LI method Confusion Matrix Bagging Algorithm

Whereas, a classification accuracy of 76.5% is obtained in ensemble creation using LTP method for imputation missing data by deploying adaBoostM1 algorithm based on confusion matrix as shown in Figure 10.

A similar percentage of 76.5% classification accuracy is obtained in ensemble creation using bagging algorithm with LTP method as illustrated in Figure 11. through confusion matrix.

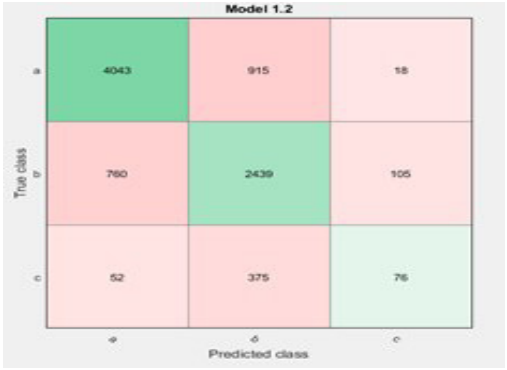


Figure 10. LTP method Confusion Matrix AdaboostM1 Algorithm

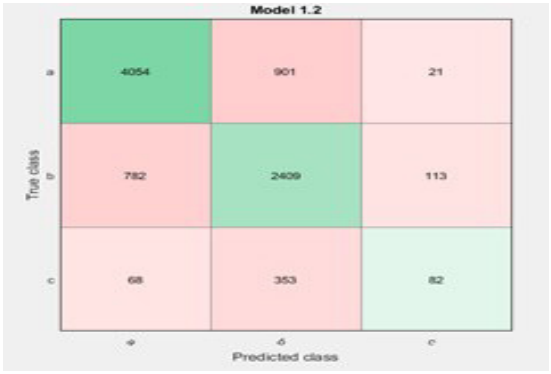


Figure 11. LTP method Confusion Matrix AdaboostM1 Algorithm

Based on the results of classification accuracy of all imputation methods, we can conclude that the best result of classification accuracy of 76.9% is obtained with SM method using adaBoostM1 and bagging algorithms. The other second best imputation method for filling missing data is MNP having classification accuracy of 76.7% with bagging and adaBoostM1 algorithms.

## 8. CONCLUSION

This study examined the effectiveness of existing SM, MNP, MDNP, LI and LTP imputation methods in terms of their error and classification accuracy in ensemble creation. There are various other effective methods proposed for dealing with missing data and tends to produce some realistic results. This research was limited to only SM, MNP, MDNP, LI and LTP imputation methods that are already implemented in SPSS and are used by various researchers resulted in useful results. However, these imputation methods performances in ensemble creation are not evaluated in the previous researches, hence, this work lead to main contribution in machine learning.

Experiment results of this research successfully identified that SM method produced lowest MAE comparing to other imputation methods in ensemble creation. Further, ensemble creation with SM method resulted in better classification accuracy compare to other methods using bagging and boosting algorithms in our research. Importantly, it is noticeable that percentage of performance accuracy margin among the imputation methods is not that high, however, SM method comparatively possessed better imputation results for our experiments.

This research is limited to small data set i.e. 8783 observations. For future work, this research work has few considerations, firstly, this research could be extended to larger data set. Secondly, further work is required for the validity of SM method results by considering various pattern of missing data. In the literature, various patterns of missing data have been used for imputation of missing data and results were obtained successfully. Thirdly, how each of these existing imputation methods in this study influences the performance of various classifiers in ensemble creation, further research on this task is also essential. Fourthly, this research could be further extended by widening the numbers of performance indicators for these five imputation methods.

## ACKNOWLEDGEMENTS

The authors would like to thanks Sreenivas Sremath Tirumala for his technical support.

## REFERENCES


[1] Lee, S.hyun. & Kim Mi Na, (2008) "This is my paper", ABC Transactions on ECE, Vol. 10, No. 5, pp120-122.

[2] Gizem, Aksahya & Ayese, Ozcan (2009) Coomunications & Networks, Network Books, ABC Publishers.

[3] Alberini, A., Bigano, A., Post, J., & Lanzi, E. (2016). Approaches and issues in valuing the costs of inaction of air pollution on human health. OECD Environment Working Papers (108), 01.

[4] Ali, S., & Tirumala, S. (2016). Performance analysis of svm ensemble methods for air pollution data. In Proceedings of the 8th international conference on signal processing systems (pp. 212–216).

[5] Ali, S., Tirumala, S. S., & Sarrafzadeh, A. (2014, Dec). Svm aggregation modelling for spatio-temporal air pollution analysis. In 17th ieee international multi topic conference 2014 (p. 249-254).

[6] Amsallem, D., & Farhat, C. (2008). Interpolation method for adapting reduced-order models and application to aeroelasticity. AIAA journal, 46(7), 1803-1813.

[7] Banjar, H., Ranasinghe, D., Brown, F., Adelson, D., Kroger, T., Leclercq, T., ... Chaudhri, N. (2017). Modelling predictors of molecular response to frontline imatinib for patients with chronic myeloid leukaemia. PloS one, 12(1), e0168947.

[8] Batista, G. E., & Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. Applied artificial intelligence, 17(5-6), 519-533.

[9] Chang, G., & Ge, T. (2011). Comparison of missing data imputation methods for traffic flow. In Transportation, mechanical, and electrical engineering (tmee), 2011 international conference on (p. 639-642).

[10] Cokluk, O., & Kayri, M. (2011). The effects of methods of imputation for missing values on the validity and reliability of scales. Educational Sciences: Theory and Practice, 11(1), 303-309.

[11] De Resende, D. C., de Santana, A. L., & Lobato, F. M. F. (2016). Time series impu-tation using genetic programming and lagrange interpolation. In Intelligent systems (bracis), 2016 5th brazilian conference on (pp. 169–174).

[12] Engels, J. M., & Diehr, P. (2003). Imputation of missing longitudinal data: a comparison of methods. Journal of clinical epidemiology, 56(10), 968-976.

[13] Eriksson, A., & van den Hengel, A. (2012, Sept). Efficient computation of robust weighted low-rank matrix approximations using the l1 norm. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(9), 1681-1690.

[14] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The kdd process for extracting useful knowledge from volumes of data. Communications of the ACM, 39(11), 27–34.

[15] Hirabayashi, S., & Kroll, C. N. (2017). Single imputation method of missing air quality data for i-tree eco analyses in the conterminous united states.

[16] Jiang, G., Tam, C. H., Luk, A. O., Kong, A. P., So, W. Y., Chan, J. C., ... Fan, X. (2016). Variable selection and prediction of clinical outcome with multiply-imputed data via bayesian model averaging. In Bioinformatics and biomedicine (bibm), 2016 ieee international conference on (p. 727-730).

[17] Jordanov, I., & Petrov, N. (2014, July). Sets with incomplete and missing data 2014; nn radar signal classification. In 2014 international joint conference on neural networks (ijcnn) (p. 218-224).

[18] Khunsongkiet, P., & Boonchieng, E. (2016, Dec). Converting air quality monitoring low cost sensor data to digital value via mobile interface. In 2016 9th biomedical engineering international conference (bmeicon) (p. 1-5).

[19] Li, Y., Li, Z., & Li, L. (2014, Feb). Missing traffic data: comparison of imputation methods. IET Intelligent Transport Systems, 8(1), 51-57.

[20] Lin, Z., Xu, C., & Zha, H. (2017). Robust matrix factorization by majorization minimization. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[21] Masiol, M., Squizzato, S., Formenton, G., Harrison, R. M., & Agostinelli, C. (2017). Air quality across a european hotspot: Spatial gradients, seasonality, diurnal cycles and trends in the veneto region, ne italy. Science of The Total Environment, 576, 210–224.

[22] Oehmcke, S., Zielinski, O., & Kramer, O. (2016, July). knn ensembles with penalized dtw for multivariate time series imputation. In 2016 international joint conference on neural networks (ijcnn) (p. 2774-2781).

[23] Pampaka, M., Hutcheson, G., & Williams, J. (2016). Handling missing data: analysis of a challenging data set using multiple imputation. International Journal of Research & Method in Education, 39(1), 19-37.

[24] Pattanodom, M., Iam-On, N., & Boongoen, T. (2016, Jan). Clustering data with the presence of missing values by ensemble approach. In 2016 second asian conference on defence technology (acdt) (p. 151-156).

[25] Rahman, A., Smith, D. V., & Timms, G. (2014, April). A novel machine learning approach toward quality assessment of sensor data. IEEE Sensors Journal, 14(4), 1035-1047.

[26] Razak, N. A., Zubairi, Y. Z., & Yunus, R. M. (2014). Imputing missing values in modelling the pm10 concentrations. Sains Malaysiana, 43(10), 1599-1607.

[27] Thirukumaran, S., & Sumathi, A. (2016, Jan). Improving accuracy rate of imputation of missing data using classifier methods. In 2016 10th international conference on intelligent systems and control (isco) (p. 1-7).

[28] Twala, B. (2005). Effective techniques for handling incomplete data using decision trees (Unpublished doctoral dissertation). Open University.

[29] Wahl, S., Boulesteix, A.-L., Zierer, A., Thorand, B., & van de Wiel, M. A. (2016). Assessment of predictive performance in incomplete data by combining internal validation and multiple imputation. BMC Medical Research Methodology, 16(1), 144.

[30] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data mining: Practical machine learning tools and techniques. Morgan Kaufmann.

[31] Wu, X., Zhu, X., Wu, G.-Q., & Ding, W. (2014). Data mining with big data. ieee transactions on knowledge and data engineering, 26(1), 97-107.

[32] Xu, Y., Du, P., & Wang, J. (2017). Research and application of a hybrid model based on dynamic fuzzy synthetic evaluation for establishing air quality forecasting and early warning system: A case study in china. Environmental Pollution.

[33] Yin, X., Levy, D., Willinger, C., Adourian, A., & Larson, M. G. (2016). Multiple imputation and analysis for high-dimensional incomplete proteomics data. Statistics in medicine, 35(8), 1315–1326.

[34] Zakaria, N. A., & NOOR, N. M. (2014). Imputation methods for filling missing data in urban air polllution data for malaysia. Urbanism. Architecture. Constructions/Urbanism. Arhitectura. Constructii, 9(2).

[35] Zhan, Y., Luo, Y., Deng, X., Chen, H., Grieneisen, M. L., Shen, X., ... Zhang, M. (2017). Spatiotemporal prediction of continuous daily pm 2.5 concentrations across china using a spatially explicit machine learning algorithm. Atmospheric Environment.

[36] Zhou, H., Zhang, D., Xie, K., & Chen, Y. (2015, Dec). Spatio-temporal tensor completion for imputing missing internet traffic data. In 2015 ieee 34th international performance computing and communications conference (ipccc) (p. 1-7).

[37] Breiman, L. (1996). Bagging predictors. Machine learning, 24(2), 123–140.

[38] Diez, J. J. R., & Gonzalez, C. J. A. (2000). Applying boosting to similarity literals for´ time series classification. In Multiple classifier systems (pp. 210–219). Springer.

[39] Xu, J., & Li, H. (2007). Adarank: a boosting algorithm for information retrieval. In Proceedings of the 30th annual international acm sigir conference on research and development in information retrieval (pp. 391–398).

## AUTHORS

Sh Shahid obtained his Master in Computer Sciences from the Auckland University of Technology, Auckland in 2006. He got industry experience as an IT Analyst and Business Analyst. He is a part time lecturer in Computer Sciences Department at Unitec. He is in his final year Doctorate having thesis entitled, "SVM Aggregation modeling for spatio-temporal air pollution analysis". He also hold some Microsoft and Cisco certifications i.e. MCP, MCSA, MCSE, CCNA, CCNP, CCDP. http://dmli.info/index.php/member.html

**Simon Dacey**

Dr. Simon Dacey is lecturer at Unitec Institute of Technology, Auckland, New Zealand and worked in software development for 17 years on applications as diverse as a firing control system for anti-submarine warfare and a database application for the management of wetland resources in Indonesia. He obtained his MSc in Applied Remote Sensing from the University of Cranfield, UK (1995). From 1998 to 2002 he worked as a lecturer at UCOL in Palmerston North, North Zealand. Since January 2002, he has been a full-time lecturer at the Department of Computing, Unitec Institute of Technology. His research interests include Geographical Information Systems, Remote Sensing, Digital Image Processing, Geographical Positioning Systems, and Database Management Systems.
http://www.unitec.ac.nz/about-us/contact-us/staff-directory/simon-dacey